DS6050

5/1/2025

Design & Evaluation of neural models for Temporal Anomaly Detection in Remote Sensing Data

Kiana Dane and Devon Hathaway

Motivation

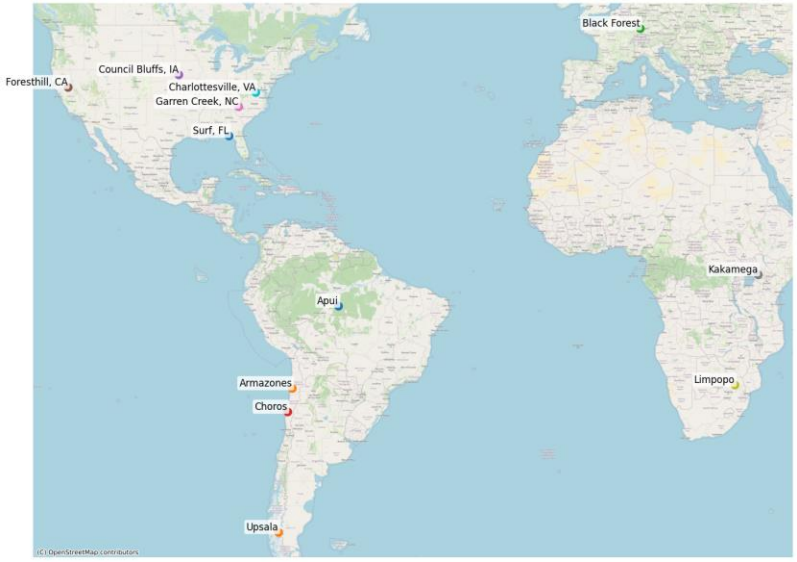
This experiment aims to contribute to the field of Earth Observation (EO). EO refers to the study of earth's surface using remote sensing technologies, such as hyperspectral imagery, typically from satellites. EO is an important field with several real-world applications, including environmental monitoring, disaster response, agricultural management, and climate change research (Li et al., 2021). Despite its potential, progress in EO research has been slow due to some significant barriers.

A primary barrier is the ambiguity to the term "anomaly." Depending on the context, anomalies can refer to many different phenomena. Some researchers might be studying drought effects, while others are looking at landslides, or illegal mining. These differences frame the crux of the issue in the field of EO, which is the difficulty in obtaining reliable labeled data. High-quality ground truth typically requires pixel-by-pixel manual annotation of imagery, a labor-intensive and often impractical process for large-scale datasets. Consequently, many EO studies must invest significant effort in creating custom-labeled datasets tailored to their specific objectives. The dataset section of this paper outlines how our team generated labeled data suitable for our anomaly detection task.

While EO as a discipline may still be considered emergent relative to more established fields, its importance is growing rapidly due to the increasing availability of high-resolution imagery and computational resources. Recent developments such as Geospatial Foundation Models (GFMs) are helping to significantly accelerate progress. These models are being trained to perform critical tasks including the rapid identification of natural disaster impacts, detection of illegal deforestation and unregulated construction, monitoring of agricultural practices, and support for global conservation initiatives (Bosilj et al., 2023).

Dataset

Images used to train and test the models were acquired from planet.com, an open-source repository that provides researchers with up-to-date high resolution satellite images for EO tasks. The dataset primarily consists of imagery captured by Planet's SuperDove satellites, which deliver 3-meter spatial resolution across eight multispectral bands: coastal blue, blue, green I, green II, yellow, red, red-edge, and near-infrared (Planet Labs, 2020). SuperDove satellites were first launched in 2020 and have been collecting imagery on a near-daily basis since their deployment. All imagery used in this study was collected between April 2020 and March 2025.



A total of 12 sites from around the world (Figure 1) that ranged in size from approximately 10km^2 to 20km^2 were chosen for this project. We purposefully chose sites at a variety of different latitudes so that we could show our model lots of seasonal variations that could help it generalize. Several sites were chosen because of known anomalies, while others were chosen as control sites. Table 1 shows more detailed information about each site.

Figure 1. Locations of each site on a world map.

Site Name	Location	Size (km$^{2)}$	Justification
Apui	Brazil	17.86	Rainforest Deforestation
Armazones	Chile	4.99	Construction of Observatory
Black Forest	Germany	15.75	Invasive Beetle Causing Mass Tree Die-Off
Choros	Chile	14.59	Rainforest, Control
Council Bluffs	Iowa, USA	10.42	Cropland, Control
Foresthill	California, USA	9.52	Forest Fire
Garren Creek	North Carolina, USA	13.20	Landslides
Kakamega	Kenya	18.76	Rainforest, Control
Limpopo	South Africa	10.11	Construction
North Pointe	Charlottesville, VA	7.76	Construction of Subdivision
Surf	Florida, USA	18.78	Coastal Region, Control
Upsala	Chile	16.42	Melting Glacier

Table 1. Sites used in testing and training for models

Images were down sampled so that each pixel contained 30m^2 of geospatial data, rather than 3m^2. This choice reduced the computing power needed to employ the deep learning approaches and it was also found that the Prithvi-EO-2.0 algorithm was more accurate when working with down sampled pixels (Szwarcman, D et al. 2024).

Our team did not have access to pre-labeled images and were therefore responsible for creating our own training data. We employed a mixture of labeling by hand and automatically. We plotted pixel reflectance values for red, green, blue and near infrared over time for a single pixel and determined if it was anomalous or normal depending on the seasonal changes throughout the five years of available data (Figure 2). We repeated this process of labeling one pixel at a time until approximately 3-5% of overall pixels had been labeled. We then automatically classified the rest by giving pixels the same label as the already labeled pixel with the shortest Euclidean distance and repeated this for all 12 sites in the dataset. This resulted in highly accurate, while not perfectly labeled training and testing data. The output of this work was two tif images: one 8-banded image of the site, and a single banded mask that included the labeled pixels for the site.

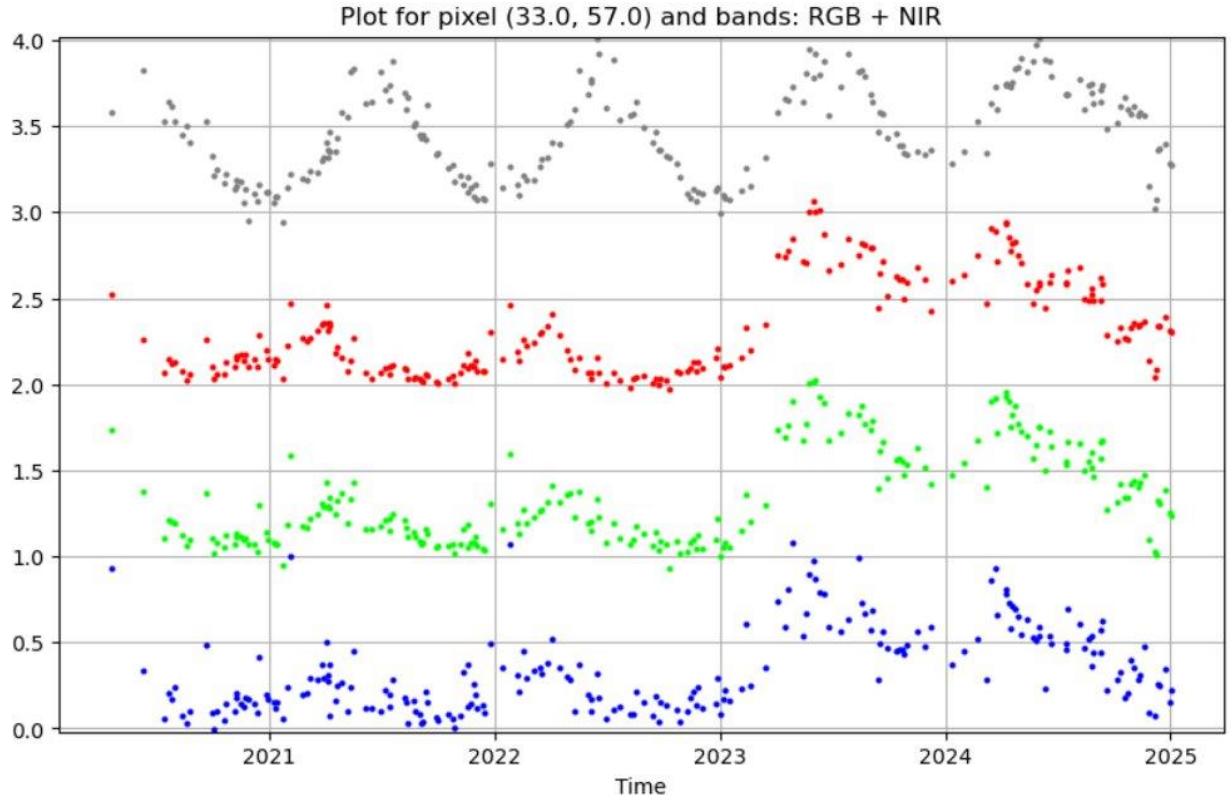


Figure 2. Example of a pixel reflectance plot over time. This plot shows an example of an anomalous pixel.

Related Work

One goal of this work is to adapt the methodology introduced by Szwarcman et al., a collaboration between IBM and NASA that resulted in the development of Prithvi-EO-2.0, a pretrained GFM based on a Vision Transformer (ViT) architecture. This model was trained on a vast dataset comprised of 4.2 million global time series samples from NASA's Harmonized Landsat and Sentinel-2 (HLS) archives at 30m² spatial resolution. With approximately 300 million trainable parameters, Prithvi-EO-2.0 has demonstrated strong performance across a range of EO tasks; however, it has not yet been applied to anomaly detection in a multi-temporal setting. The use case presented in their paper that most closely aligns with our objectives is the multi-temporal crop classification task, where the model was trained and evaluated across 13 crop classes over 80 epochs, achieving an F1 score of 84.4% using the full training dataset. Given the model's demonstrated capacity to learn spatiotemporal representations, we believe the Prithvi-EO-2.0 model is a good candidate for adaptation to the underexplored task of anomaly detection in hyperspectral satellite imagery.

Technical Approach

Prithvi-EO-2.0 foundational model approach

To ensure compatibility with the Prithvi-EO-2.0 model, several preprocessing steps were required. Our original dataset consisted of eight spectral bands; however, only four bands (red, blue, green, and near infrared) aligned with the inputs used during Prithvi's pretraining. The remaining bands were removed to maintain consistency. Additionally, the standard Prithvi model expects six input bands; therefore, we modified the model architecture to accept four-channel inputs. Prithvi-EO-2.0 expects square input patches of a fixed size. To satisfy this requirement, we divided each satellite image and its corresponding labeled mask into 56×56 pixel tiles.

To use the model for temporal analysis, we prepared the data by stacking multiple timestamps into a single input array. Initially, we attempted to include one image per month over five years (resulting in an input with 60 temporal steps). However, the resulting memory requirements exceeded the capacity of both local and Rivanna high-performance computing environments. As an alternative, we stacked images acquired during the same month across five consecutive years, yielding inputs with five timesteps. While this approach reduced the temporal resolution and solved any memory issues, it may negatively impact overall model performance. To evaluate model generalization, we divided the dataset into four distinct train/test splits, ensuring that each site appeared exactly once in the test set. Strict data separation was maintained to prevent data leakage between training and evaluation.

The Prithvi-EO-2.0 GFM is based on a ViT architecture and is pretrained using a masked autoencoder (MAE) strategy. Unlike standard 2D MAEs, Prithvi adapts the architecture to the spatiotemporal nature of EO data by introducing 3D patch embeddings and 3D positional encodings that account for the spatial

(height and width) and temporal (time) dimensions of satellite image sequences. Additionally, the model incorporates auxiliary metadata (latitude and longitude) through independent 2D sinusoidal embeddings, which are added as biases to the embedded tokens. To improve robustness, metadata dropout is applied during pretraining, randomly omitting temporal or location information.

During fine-tuning, we employed a weighted cross-entropy loss function, with class weights inversely proportional to the anomaly rate in the training set (excluding the validation and testing data). An AdamW optimizer was used for optimization. Random horizontal and vertical flips were applied as data augmentation strategies during training. Due to the lack of early stopping mechanisms in the current Prithvi fine-tuning framework, we manually experimented with different epoch counts. Based on validation performance, eight training epochs were selected as the optimal point to balance learning progress and overfitting.

Transformer Encoder with Temporal Attention

To explore an entirely different approach, we also developed a custom Transformer Encoder with Temporal Attention, designed specifically to model complex temporal dynamics across high-dimensional Earth observation data. This architecture captures both short- and long-range dependencies across time while being flexible enough to process varying-length sequences across diverse geographies.

In parallel, we benchmarked our model against the Prithvi Foundation Model—a state-of-the-art transformer pre-trained on planetary-scale satellite imagery—as a strong baseline for transfer learning.

Our Temporal Transformer consists of the following components:

- Input Projection: A linear layer maps each multi-spectral input vector to a fixed d_model dimension.
- Positional Encoding: We use learnable positional encodings to retain temporal ordering of satellite observations.
- Transformer Encoder Stack: Multiple layers of multi-head self-attention allow each timestamp to attend to all others, enabling context-aware anomaly detection.
- Temporal Pooling: After encoding, we apply mean pooling across time steps to create a fixed-size sequence embedding.
- Classification Head: A fully connected feedforward network predicts whether a temporal sequence contains an anomaly.

Experiments

When evaluating model performance, we aimed to assess not only the overall classification accuracy but also the model's ability to correctly detect rare anomalies, which is critical in anomaly detection tasks. Thus, in addition to accuracy, we report precision, recall, F1 score, false positive rate (FPR), and false

negative rate (FNR). Anomalies in the wild tend to be rare, so these metrics provide a more nuanced view of performance, where traditional accuracy alone may be misleading. In particular, precision and recall are important indicators of how well the model identifies true anomalies while avoiding false detections.

Prithvi-EO-2.0

Prithvi-EO-2.0 was fine-tuned on our labeled anomaly dataset using four-fold train/test splits. Each site appeared exactly once in the test set across the splits. The results presented below are averages across these four splits.

Metric	Result (%)
Overall Accuracy	73.1%
Precision	26.4%
Recall	32.9%
F1 Score	21.9%
FPR	21.9%
FNR	67.1%

Table 2. Average reporting metrics for Prithvi-EO-2.0 model across all data

The overall classification accuracy of 73.1% suggests that the model was able to correctly classify most pixels, but the relatively low precision (26.4%) and recall (32.9%) indicate challenges in correctly identifying anomalous regions. The high FNR (67.1%) highlights that many anomalies were missed, which is consistent with the difficulty of detecting rare events in highly imbalanced datasets. Meanwhile, the FPR of 21.9% shows that a moderate proportion of normal pixels were misclassified as anomalous.

One of the primary limitations that likely contributed to the modest performance of the model was the size and scope of the available dataset. Due to constraints in data access through planet.com, we were limited in the number of usable images we could collect across different sites and time periods. Furthermore, the SuperDove satellite imagery required for our study was only available starting in 2020, restricting the temporal depth of our dataset to just five years. This relatively short observation window, combined with a limited number of spatial sites, likely hindered the model's ability to learn diverse patterns of normal and anomalous behavior. In most cases, anomalies tend to be much rarer and more diverse than non-anomalies, and therefore a more varied dataset is critical for achieving higher precision and recall.

Transformer Encoder with Temporal Attn.

While Prithvi benefited from strong pretraining and general feature representations, it struggled to model site-specific temporal deviations and exhibited higher false positive rates. Our custom transformer, trained end-to-end on the task-specific anomaly labels, demonstrated better temporal generalization and interpretability when visualized across site timelines.

However, the Prithvi model still offers advantages in scenarios with limited labels, thanks to its strong pretraining. Future directions include integrating the temporal transformer within a Prithvi-style encoder or using it for fine-tuning Prithvi representations to achieve the best of both worlds.

Metric	Result (%)
Overall Accuracy	43.07%
Precision	7.9%
Recall	72.6%
F1 Score	13.87%
FPR	73.1%
FNR	17.6%

Table 3. Average reporting metrics across all sites for Transformer Encoder with Temporal Attn

Conclusion

Overall, the Prithvi-EO-2.0 model outperformed our transformer encoder model. This is likely due to the fact that the Prithvi model is pretrained on millions of real images from around the world and is therefore able to generalize across a variety of sites much better. This ability to generalize was especially important given that our biggest limitation was a small, relatively narrow training dataset.

If we were to continue our work with the Prithvi model, we would start by collecting more data for training and testing. We would also work on a solution for the memory issue we experienced and attempt to successfully run the algorithm with all 60 timesteps in a single array, rather than breaking it up by month. Finally, we would also consider creating more classes to identify different types of anomalies (for example, differentiating between a forest fire scar, new construction, or deforestation).

The field of EO is incredibly important to efforts in crisis response, conservation and much more, and should therefore continue to be studied. Our work highlights both the potential and the challenges of adapting geospatial foundation models for anomaly detection.

References

1. Li, X., Zhang, C., & Ge, Y. (2021). "Remote sensing image classification: A review." *Remote Sensing*, 13(17), 3411. [DOI:10.3390/rs13173411]
2. Bosilj, P., Veličković, P., Edvardsen, D., & Gadde, R. (2023). "Foundation models for Earth Observation: Challenges and Opportunities." *arXiv preprint* arXiv:2307.05281.
3. Planet Labs PBC (2020). *SuperDove Technical Specifications.* https://assets.planet.com/docs/Planet-Scope-Product-Specs-2020.pdf
4. Szwarcman, D., Roy, S., Fraccaro, P., et al. (2024). *Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications.* arXiv:2412.02732v2.